Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 41-48 recibido 04-05-12 revisado 03-06-12 aceptado 06-06-12

Análisis de técnicas PLN de expansión de consulta aplicadas a la tarea de la recuperación de información geográfica*

Analysis of NLP techniques of query expansion applied to the Geographical Information Retrieval task

José M. Perea-Ortega **Miguel Á. García-Cumbreras**
L. Alfonso Ureña-López **Arturo Montejo-Ráez**

Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
{jmperea,magc,laurena,amontejo}@ujaen.es

Resumen: En este trabajo, proponemos diferentes técnicas relacionadas con el Procesamiento del Lenguaje Natural (PLN) para reformular las consultas geográficas lanzadas a un sistema GIR. Estás técnicas consistirán en la modificación y/o expansión de las dos partes normalmente reconocidas en una consulta geográfica: la parte temática y la parte geográfica. Hemos evaluado cada una de las reformulaciones propuestas utilizando un marco de experimentación para evaluar sistemas GIR como GeoCLEF. Los resultados obtenidos demuestran que todas las reformulaciones de consulta propuestas recuperaron documentos relevantes que no fueron recuperados utilizando la consulta original, por lo que estas estrategias se pueden considerar de utilidad a la hora de trabajar con sistemas GIR.
Palabras clave: Reformulación de consulta geográfica, recuperación de información geográfica, GeoCLEF

Abstract: In this paper, we propose different Natural Language Processing (NLP) techniques of query reformulation related to the modification and/or expansion of the thematic and geographic parts that are usually identified in a geographic query. We have evaluated each of the reformulations proposed using GeoCLEF as an evaluation framework for GIR systems. The results obtained show that all proposed query reformulations retrieved new relevant documents that were not retrieved using the original query.
Keywords: Geographic query reformulation, Geographical Information Retrieval, GeoCLEF

1. Introducción

En el campo de la Recuperación de Información (*Information Retrieval*, IR) (Baeza-Yates y Ribeiro-Neto, 1999), al enfoque basado en la modificación de la consulta del usuario para mejorar la calidad de los resultados de la búsqueda se le conoce como reformulación de consulta. El objetivo de dicho proceso es satisfacer la necesidad de información de los usuarios, normalmente mejorando la calidad y la cobertura de los resultados obtenidos mediante la consulta original. Esta característica la soportan algunos motores de búsqueda de forma explícita, sugiriendo búsquedas similares a la consulta inicial del usuario. Por otra parte, algunos motores de búsqueda también consiguen reformular la consulta de forma implícita, es decir, mediante la expansión de la consulta original con términos relacionados con sus palabras clave, por ejemplo.

La recuperación de información geográfica (*Geographical Information Retrieval*, GIR) es un área de investigación activa y en crecimiento que se centra en la recuperación de documentos textuales de acuerdo a un criterio geográfico de relevancia. Por esta razón, la GIR se puede considerar una extensión de la IR. En concreto, la GIR se ocupa de mejorar

* Este trabajo ha sido cofinanciado por la Comisión Europea bajo el Séptimo Programa Marco (FP7-2007-2013) a través del proyecto FIRST (FP7-287607), por el Fondo Europeo de Desarrollo Regional (FEDER) con el proyecto TEXT-COOL 2.0 (TIN2009-13391-C04-02) y por el proyecto local Geocaching Urbano (RFC/IEG2010)

la calidad de la recuperación específica de la información geográfica, centrándose en el acceso a documentos no estructurados (Jones y Purves, 2008; Larson, 1996). La comunidad IR ha sido principalmente responsable de la investigación en el campo de la GIR, en lugar de la comunidad relacionada con los Sistemas de Información Geográfica (SIG). El tipo de consulta en un motor IR se basa generalmente en lenguaje natural, en contraste con el enfoque común más formal de los SIG, en los que cada objeto geográfico se recupera de una base de datos estructurada. En un sistema GIR, una consulta geográfica puede ser estructurada como una tripleta <*tema*><*relación espacial*><*localización*>, donde <*tema*> es el foco principal de la consulta, <*localización*> representa el ámbito geográfico de la consulta y <*relación espacial*> determina la relación entre el foco o tema y el ámbito geográfico. Por ejemplo, la tripleta para la consulta geográfica "*airplane crashes close to Russian cities*" sería de <*airplane crashes*> <*close to*> <*Russian cities*>. En definitiva, para una búsqueda como "*castles in Spain*", un buen sistema GIR debería devolver no sólo los documentos que contienen la palabra "*castles*", sino también aquellos que tengan alguna entidad geográfica relacionada con España.

Dado que un sistema GIR puede ser visto o tratado como un motor de búsqueda tradicional (los resultados para una consulta se muestran como una lista de documentos ordenada de mayor a menor relevancia), es importante prestar atención a la búsqueda de métodos eficaces para reformular la consulta de usuario. Estos métodos pueden tener en cuenta tanto características léxico-sintácticas como aspectos geográficos. De esta manera, los resultados de la búsqueda mejorarán su calidad y su cobertura. El objetivo de este trabajo es evaluar y analizar el comportamiento de varias reformulaciones de consulta geográfica propuestas para la tarea GIR, teniendo en cuenta que un sistema GIR puede funcionar como un sistema IR. Para llevar a cabo esta evaluación, se ha utilizado el marco de trabajo más importante en este contexto: GeoCLEF[1] (Gey et al., 2005; Mandl et al., 2008).

El resto del artículo se estructura de la siguiente manera: en la Sección 2 se exponen los trabajos más importantes relacionados con la reformulación de consulta geográfica en general; en la Sección 3 se describe el sistema GIR utilizado para los experimentos; en la Sección 4 se describe brevemente el marco de evaluación; en la Sección 5 se presentan los experimentos realizados y un análisis de los resultados obtenidos; por último, en la Sección 6, se exponen las conclusiones y el trabajo futuro.

2. Trabajo relacionado

Jansen, Booth, y Spink (2009) definen el concepto de reformulación de consulta como el proceso de alteración de la consulta para mejorar el funcionamiento de la búsqueda o recuperación de información. En algunas ocasiones, los motores de búsqueda aplican este método utilizando la técnica conocida como "retroalimentación por relevancia" (*relevance feedback*). Esta técnica se puede aplicar, por un lado, permitiendo que los usuarios decidan si un documento recuperado es relevante o no, generando, a partir de su elección, reformulaciones de la consulta original automáticamente. Por otro lado, esta reformulación también se puede llevar a cabo de manera automática analizando los n primeros documentos recuperados sin la intervención del usuario, teniendo en cuenta estadísticas relacionadas con cada término del documento. No obstante, se ha estudiado que los usuarios pocas veces utilizan la retroalimentación por relevancia (Spink, Jansen, y Ozmultu, 2000) y normalmente reescriben su consulta de forma manual (Anick, 2003).

Este trabajo está relacionado específicamente con consultas de tipo geográfico. Según Gravano, Hatzivassiloglou, y Lichtenstein (2003), los motores de búsqueda actuales son criticados debido a su ignorancia por no considerar restricciones geográficas en las consultas de los usuarios y, por tanto, eso provoca que recuperen menos documentos relevantes. Este problema podría ser atribuido a la manera en la que los motores de búsqueda tradicionales manejan las consultas en general, ya que normalmente adoptan un enfoque basado en la coincidencia con las palabras clave de la consulta (*keywords matching*), sin tener en cuenta el ámbito espacial de los términos geográficos.

Varios autores han realizado estudios sobre las búsquedas realizadas por usuarios cuando utilizan consultas geográficas en mo-

[1] `http://ir.shef.ac.uk/geoclef/`

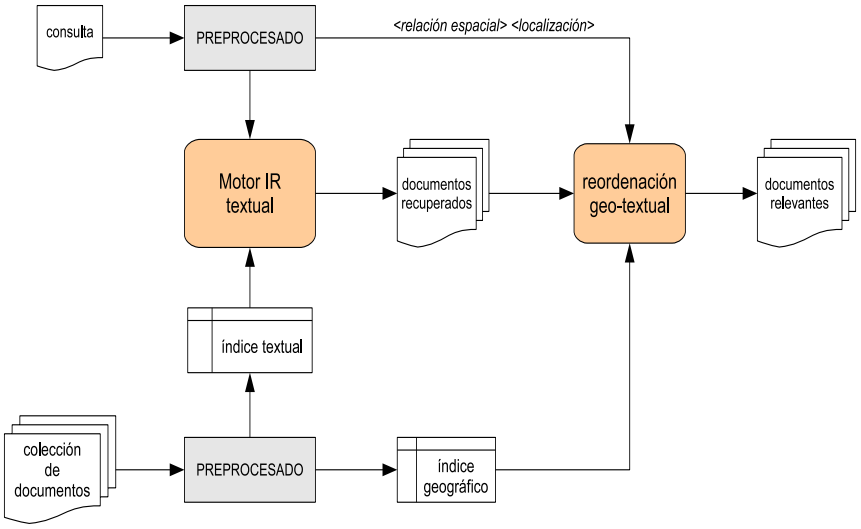

Figura 1: Arquitectura general del sistema SINAI-GIR

tores de búsqueda (Sanderson y Kohler, 2004; Gan et al., 2008; Jones et al., 2008). Una de las principales conclusiones de estos estudios es que la estructura de dichas consultas consistía normalmente en dos partes, una temática y otra geográfica, conteniendo esta última términos espaciales o direccionales. Desde un punto de vista geográfico, Kohler (2003) proporciona un estudio sobre geo-reformulación de consultas. Concluye que la adición de más términos geográficos en la consulta es una técnica utilizada comúnmente para diferenciar lugares que comparten el mismo nombre. A esta técnica se le conoce como expansión de consulta utilizando entidades geográficas.

En la literatura podemos encontrar diferentes trabajos que han abordado la expansión de consulta geográfica. Cardoso y Silva (2007) presentan un enfoque basado en el uso de tipos de característica (isla, ciudad, montaña), reajustando la estrategia de expansión de acuerdo a la semántica de la consulta. Fu, Jones, y Abdelmoty (2005) proponen un método de expansión basado en ontología que soporta la recuperación de documentos que son considerados geográficamente relevantes. Consiguen mejorar los resultados de búsqueda cuando la consulta contiene una relación espacial difusa, mostrando que el método propuesto funciona eficientemente utilizando ontologías realistas en un entorno de búsqueda espacial distribuido. Buscaldi, Rosso, y Sanchis Arnal (2005) utilizan WordNet[2] durante la fase de indexación, añadiendo los sinónimos y los holónimos de las entidades geográficas detectadas en el índice de términos de cada documento, demostrando la efectividad de dicho método. Por último, Stokes et al. (2008) concluyen que se obtiene una ganancia significativa en un sistema GIR cuando todos los conceptos (no sólo los geográficos) son expandidos.

3. *La arquitectura SINAI-GIR*

En esta sección se describe un ejemplo de sistema GIR. Específicamente, hemos utilizado nuestro propio sistema GIR llamado SINAI-GIR (Perea-Ortega et al., 2008b). Al igual que ocurre en los sistemas IR tradicionales, en un sistema GIR podemos diferenciar tres etapas: procesamiento de la colección de documentos y las consultas, indexación y búsqueda textual-geográfica y, finalmente, reordenación de los documentos recuperados utilizando una fórmula de relevancia particular que combina la similitud geográfica y textual entre la consulta y el documento recuperado. El sistema GIR utilizado en este trabajo sigue un enfoque similar, como se puede observar en la Figura 1.

Durante el proceso de funcionamiento de un sistema GIR, cada consulta es preprocesada y analizada, identificando el ámbito geográfico y la relación espacial que pudiera contener. Por otro lado, la colección de

[2] `http://wordnet.princeton.edu/`

documentos también es preprocesada, detectando todas las entidades geográficas y generando un índice geográfico con ellas. Durante esa fase, cada consulta preprocesada (incluyendo sus entidades geográficas) es lanzada contra el motor de búsqueda. Finalmente, los documentos recuperados son filtrados y reordenados, colocando en las últimas posiciones aquellos documentos cuyo ámbito geográfico no coincide con el detectado en la consulta. Por el contrario, aquellos documentos cuya similitud con el ámbito geográfico de la consulta es mayor, son colocados en las primeras posiciones de la lista resultado.

Con respecto al procesamiento de la consulta, éste se basa principalmente en el reconocimiento de las entidades geográficas. Además, también implica la especificación de la tripleta explicada en la Sección 1, que será utilizada más tarde durante el proceso de filtrado y reordenación. Para detectar dicha tripleta, hemos utilizado un etiquetador de la categoría léxica de las palabras (*Part Of Speech*, *POS tagger*) como TreeTagger[3], teniendo en cuenta además algunas reglas sintácticas como *preposición + nombre propio*, por ejemplo. Las palabras vacías fueron eliminadas y se aplicó un extractor de raíces como Snowball[4] a cada palabra, excepto para las entidades geográficas, que mantuvieron su forma original.

Durante el proceso de recuperación, se obtuvieron 1.000 documentos para cada consulta, utilizando Terrier[5] como motor de recuperación. Según un estudio previo realizado por Perea-Ortega et al. (2008a), se demostró que Terrier es una de las herramientas más utilizadas para aplicaciones IR en general y en sistemas GIR en particular, obteniendo resultados prometedores. El esquema de pesado utilizado ha sido *inL2*, el cúal está implementado por defecto en Terrier. Este esquema aplica el modelo de la frecuencia inversa de documento (*Inverse Document Frequency*, IDF) para la aleatoriedad, la sucesión de Laplace para la normalización en primer lugar y normalización 2 para la normalización de la frecuencia de cada término (Amati, 2003). Por último, cabe señalar que, aunque los sistemas GIR normalmente aplican un proceso de geo-reordenación después del módulo IR, en este trabajo no es necesario utilizarlo porque estamos interesados en evaluar la precisión y cobertura de cada reformulación de consulta propuesta desde el punto de vista de la recuperación de información.

4. GeoCLEF como marco de evaluación

Para evaluar las reformulaciones propuestas hemos utilizado el marco de experimentación GeoCLEF (Gey et al., 2005; Mandl et al., 2008), un foro de evaluación para sistemas GIR celebrado entre los años 2005 y 2008 bajo el marco de las conferencias CLEF[6]. GeoCLEF proporciona una colección de 169.477 documentos que consisten en noticias extraídas del periódico británico *Glasgow Herald* (1995) y del periódico américano *Los Angeles Times* (1994), representando una amplia variedad de regiones geográficas y lugares. Por otro lado, se proporcionaron un total de 100 consultas textuales o *topics* (25 consultas por año). Las consultas están compuestas por tres campos principales: título (T), descripción (D) y narrativa (N). Para los experimentos llevados a cabo en este trabajo, sólo hemos tenido en cuenta el campo título, ya que representa de forma similar la manera en la que un usuario lanzaría una consulta geográfica a un motor de búsqueda. Algunos ejemplos de consultas GeoCLEF son: "*vegetable exporters of Europe*", "*forest fires in north of Portugal*", "*airplane crashes close to Russian cities*" or "*natural disasters in the Western USA*".

Con respecto a las medidas de evaluación utilizadas, los resultados han sido evaluados haciendo uso de los juicios de relevancia proporcionados por los organizadores de GeoCLEF y del método de evaluación TREC[7]. La evaluación se ha realizado utilizando las medidas típicas de evaluación en recuperación de información: precisión media (*Mean Average Precision*, MAP), cobertura (*Recall*, R) y precisión en n (P@n).

5. Experimentos y resultados

Tal y como se ha comentado previamente, en este trabajo se analizan varias reformulacio-

[3]TreeTagger v.3.2 para Linux. Disponible en `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

[4]Disponible en `http://snowball.tartarus.org`

[5]Versión 2.2.1, disponible en `http://terrier.org`

[6]Cross Language Evaluation Forum `http://www.clef-initiative.eu/`

[7]*trec_eval* es un programa para evaluar resultados TREC utilizando los procedimientos estándar del NIST `http://trec.nist.gov/trec_eval/`

Reformulación	Texto de la consulta
original	`visit American presid Germany`
QR1	`visit American presid`
QR2	`visit American presid visit American presid Germany`
QR3	`#and(#or(visit meet stay ) American presid Germany)`
QR4	`#and(visit American presid #or(Germany #3(Federal Republic of Germany) Deutschland FRG ) )`
QR5	`#and(visit American presid #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen ) )`
QR6	`#and(#or(visit meet stay ) of the American presid ) #or(Germany Berlin Hamburg Muenchen Koeln #2(Frankfurt am Main) Essen )`

Tabla 1: Ejemplo de reformulaciones generadas para la consulta "*Visits of the American president to Germany*"

nes de consultas para la tarea GIR. Para ello se han utilizado las dos partes principales de una consulta geográfica: la parte temática y el ámbito geográfico detectado. El objetivo de estas reformulaciones es mejorar el proceso de recuperación, tratando de encontrar documentos relevantes que no han sido recuperados utilizando la consulta original. A partir de la consulta original preprocesada, hemos generado los siguientes tipos de reformulación:

- QR1: el ámbito geográfico es eliminado, dejando únicamente la parte temática de la consulta original.
- QR2: la parte temática es expandida repitiendo sus palabras clave. De esta forma, tratamos de dar más importancia a la parte temática con respecto a la parte geográfica.
- QR3: la parte temática es expandida utilizando únicamente los sinónimos de las palabras clave detectadas en dicha parte. Hemos considerado como palabras clave los sustantivos, por ser aquellos que más carga semántica poseen. Se ha utilizado WordNet como recurso léxico para la obtención de los sinónimos, consultando todos los sentidos del sustantivo.
- QR4: la parte geográfica es expandida utilizando únicamente sinónimos del ámbito geográfico detectado en la consulta. Se ha utilizado GeoNames[8] como base de conocimiento geográfico.
- QR5: la parte geográfica es expandida utilizando localizaciones o lugares que coinciden con el ámbito geográfico y la relación espacial detectados en la consulta.
- QR6: tanto la parte temática como la geográfica son expandidas, combinando las reformulaciones QR3 y QR5.

La Tabla 1 muestra un ejemplo de las diferentes reformulaciones generadas para la consulta "*Visits of the American president to Germany*". Como se puede observar, en las reformulaciones QR2 y QR3 se expande únicamente la parte temática, mientras que en las reformulaciones QR4 y QR5 se expande solamente la parte geográfica. Finalmente, la reformulación QR6 puede ser considerada una combinación de expansiones utilizando ambas partes.

Los diferentes resultados obtenidos utilizando cada reformulación de consulta (*Query Reformulation*, QR), junto con el obtenido usando la consulta original, se pueden observar en la Tabla 2. En dicha tabla, se muestra la precisión media en los 5, 10 y 100 primeros documentos recuperados, la cobertura y el valor MAP para cada reformulación. Aunque ninguna de ellas consigue mejorar el MAP obtenido con la consulta original, es interesante señalar que la QR2 (la parte temática es expandida repitiendo sus palabras clave) alcanza el mejor resultado P@10 en tres de los cuatro grupos de consultas.

Llegados a este punto, nos preguntamos si las reformulaciones propuestas estaban realmente recuperando documentos relevantes que la consulta original no era capaz de recuperar. Para resolver esta duda, utilizamos los juicios de relevancia proporcionados por los

[8] `http://www.geonames.org`

Grupo de consultas	QR	P@5	P@10	P@100	R	MAP
2005	original	**0,5520**	0,4560	**0,1904**	**0,8364**	**0,3514**
	QR1	0,2640	0,2560	0,1260	0,6748	0,1638
	QR2	0,5200	**0,4920**	0,1840	0,8276	0,3353
	QR3	0,3680	0,3160	0,1400	0,7596	0,2035
	QR4	0,3120	0,2800	0,1212	0,6552	0,2242
	QR5	0,1440	0,1240	0,0772	0,5624	0,0952
	QR6	0,1600	0,1480	0,0780	0,5692	0,0942
2006	original	**0,2400**	0,1920	**0,0716**	**0,7288**	**0,2396**
	QR1	0,0560	0,0640	0,0252	0,4604	0,0615
	QR2	0,2320	**0,2040**	0,0664	0,6796	0,2314
	QR3	0,1440	0,1400	0,0604	0,7356	0,1419
	QR4	0,1920	0,1720	0,0636	0,6984	0,2064
	QR5	0,2240	0,1840	0,0612	0,6524	0,1811
	QR6	0,1840	0,1760	0,0580	0,6772	0,1486
2007	original	**0,3040**	**0,2560**	**0,1188**	**0,7156**	**0,2311**
	QR1	0,1600	0,1320	0,0796	0,4452	0,1255
	QR2	0,2640	0,2120	0,1072	0,6656	0,1871
	QR3	0,2000	0,1800	0,0884	0,6284	0,1774
	QR4	0,2160	0,2000	0,1020	0,6608	0,1687
	QR5	0,2240	0,2000	0,0928	0,6720	0,1874
	QR6	0,2240	0,2040	0,0836	0,6344	0,1763
2008	original	**0,3760**	**0,2680**	0,1104	**0,7368**	**0,2484**
	QR1	0,1760	0,1400	0,0928	0,5996	0,1301
	QR2	0,3440	**0,2680**	**0,1124**	0,7196	0,2381
	QR3	0,2960	0,2320	0,1024	0,6884	0,1972
	QR4	0,2640	0,1960	0,0924	0,6404	0,1619
	QR5	0,2720	0,2040	0,0964	0,6984	0,1906
	QR6	0,2720	0,2280	0,0948	0,7028	0,2028

Tabla 2: Resultados obtenidos para cada reformulación de consulta propuesta

Grupo de consultas	Nº total docs relev.	ORIG	QR1	QR2	QR3	QR4	QR5	QR6
2005	1028	88,33 %	2,33 %	2,04 %	1,85 %	0,68 %	1,36 %	1,85 %
2006	378	75,13 %	2,65 %	1,32 %	4,76 %	1,59 %	3,97 %	5,56 %
2007	650	83,54 %	1,08 %	1,08 %	2,77 %	5,38 %	4,62 %	5,23 %
2008	747	78,71 %	4,82 %	4,28 %	2,95 %	0,80 %	9,10 %	8,97 %
Media		81,43 %	2,72 %	2,18 %	3,08 %	2,11 %	4,76 %	5,40 %

Tabla 3: Porcentaje de documentos relevantes recuperados por cada reformulación propuesta comparados con la consulta original

organizadores de GeoCLEF y obtuvimos los resultados que se muestran en la Figura 2. El número total de documentos recuperados fue siempre 1.000. Por otro lado, según los juicios de relevancia, el número total de documentos relevantes para cada conjunto de consultas (2005, 2006, 2007 y 2008) fue 1.028, 378, 650 y 747, respectivamente. Además, el número de documentos relevantes recuperados por la consulta original fue 908, 284, 543 y 588 para el conjunto de topics 2005, 2006, 2007 y 2008, respectivamente. La Tabla 3 muestra el porcentaje de documentos relevantes recuperados por cada reformulación propuesta comparados con la consulta original.

Analizando estos resultados, podemos observar que todas las reformulaciones propuestas siempre recuperan documentos relevantes que no fueron recuperados por la consulta original. Cabe destacar el buen comportamiento en general de las reformulaciones basadas en la expansión de la parte geográfica (QR4 y

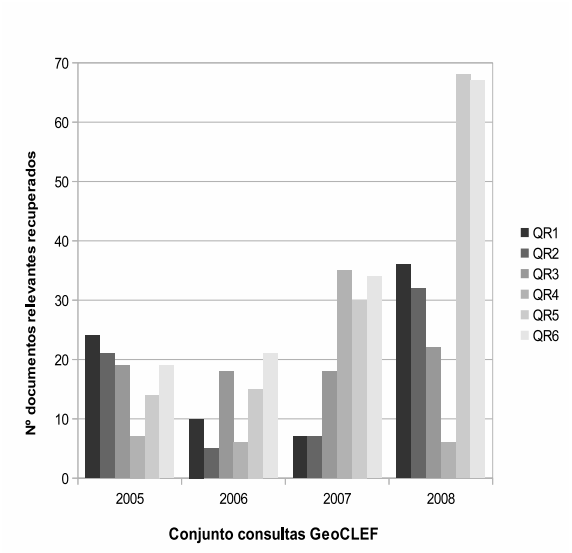

Figura 2: Comparativa del número de documentos relevantes recuperados para cada reformulación que no fueron recuperados por la consulta original

QR5). Específicamente, la QR5 alcanza una importante diferencia utilizando los topics de 2008, con un total de 68 documentos relevantes que no fueron recuperados por la consulta original, es decir, un 9,10 % del total de documentos relevantes para esos topics (ver Tabla 3). Esto significa que de los 159 (747-588) documentos relevantes no recuperados por la consulta original para esos topics, el 42,77 % de ellos fueron recuperados utilizando dicha reformulación. Otro ejemplo parecido ocurre con la QR4, que obtiene el valor más alto para los topics de 2007 consiguiendo el 5,38 % de los documentos relevantes para esas consultas. Este dato representa el 32,71 % de los documentos relevantes no recuperados por la consulta original.

Con respecto a las reformulaciones relacionadas con la expansión de la parte temática (QR2 y QR3) éstas también obtuvieron buenos resultados en general, aunque especialmente para los topics de 2005 y 2006. La QR3 consiguió un 4,76 % de los documentos relevantes para los topics de 2006. Todo esto hace que la reformulación QR6 que combina las QR3 y QR5 obtenga muy buenos resultados como se muestra para todos los conjuntos de consultas en general. Finalmente, comparando cada reformulación con el resto en cada conjunto de consultas, la reformulación QR1 obtiene el mejor valor para las de 2005, por lo que la idea de eliminar la parte geográfica en la consulta original puede ser una buena estrategia cuando la consulta se considera no geográfica.

6. Conclusiones y trabajo futuro

En este trabajo se proponen diferentes técnicas PLN de reformulación de consulta basadas en la modificación y/o expansión de las dos partes de una consulta geográfica: la parte temática y el ámbito geográfico. Se han evaluado cada una de las reformulaciones propuestas utilizando GeoCLEF como marco de evaluación para sistemas GIR. Esta evaluación se ha llevado a cabo desde el punto de vista de la recuperación de información, es decir, sin tener en cuenta nigún tipo de proceso de geo-reordenación posterior al proceso de recuperación de documentos relevantes. Los resultados obtenidos demuestran que, si bien el rendimiento no mejora respecto del caso base, todas las reformulaciones propuestas recuperaron documentos relevantes que no fueron recuperados mediante la consulta original. Esto nos lleva a pensar que, en determinados casos, es necesario realizar estas reformulaciones, si bien no lo es para todas las consultas.

Por tanto, como trabajo futuro, analizaremos los diferentes tipos de consultas geográficas para así estudiar con más profundidad en qué casos es aconsejable aplicar estas técnicas en un sistema GIR dependiendo del tipo de consulta. De este modo, aplicando una expansión selectiva, trataremos de mejorar el rendimiento general del sistema cuando se utiliza únicamente la consulta original.

Bibliografía

Amati, G. 2003. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. Ph.D. tesis, School of Computing Science, University of Glasgow.

Anick, Peter. 2003. Using terminological feedback for web search refinement: a log-based study. En *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, páginas 88–95, New York, NY, USA. ACM.

Baeza-Yates, Ricardo A. y Berthier Ribeiro-Neto. 1999. *Modern Information Retrie-*

val. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Buscaldi, Davide, Paolo Rosso, y Emilio Sanchis Arnal. 2005. Using the wordnet ontology in the geoclef geographical information retrieval task. En *CLEF*, volumen 4022 de *Lecture Notes in Computer Science*, páginas 939–946. Springer.

Cardoso, Nuno y Mário J. Silva. 2007. Query expansion through geographical feature types. En Ross Purves y Chris Jones, editores, *GIR*, páginas 55–60. ACM.

Fu, Gaihua, Christopher B. Jones, y Alia I. Abdelmoty. 2005. Ontology-based spatial query expansion in information retrieval. En *OTM Conferences (2)*, volumen 3761 de *Lecture Notes in Computer Science*, páginas 1466–1482. Springer.

Gan, Qingqing, Josh Attenberg, Alexander Markowetz, y Torsten Suel. 2008. Analysis of geographic queries in a search engine log. En *Proceedings of the first international workshop on Location and the web*, páginas 49–56, Beijing, China. ACM.

Gey, Fredric C., Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, y Vivien Petras. 2005. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. En *CLEF*, volumen 4022 de *Lecture Notes in Computer Science*, páginas 908–919. Springer.

Gravano, L., V. Hatzivassiloglou, y R. Lichtenstein. 2003. Categorizing web queries according to geographical locality. En *Proceedings of the 12th International Conference on Information and Knowledge Management*, páginas 325–333.

Jansen, Bernard J., Danielle L. Booth, y Amanda Spink. 2009. Patterns of query reformulation during web searching. *JASIST*, 60(7):1358–1371.

Jones, Christopher B. y Ross S. Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones, Rosie, Wei Vivian Zhang, Benjamin Rey, Pradhuman Jhala, y Eugene Stipp. 2008. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246.

Kohler, J. 2003. Analysing search engine queries for the use of geographic terms. Master's thesis, University of Sheffield - United King.

Larson, R. 1996. Geographic information retrieval and spatial browsing. En Smith y M. Gluck, editores, *Geographic Information Systems and Libraries: Patronsand Mapsand and Spatial Information*, páginas 81–124.

Mandl, Thomas, Paula Carvalho, Giorgio Maria Di Nunzio, Fredric C. Gey, Ray R. Larson, Diana Santos, y Christa Womser-Hacker. 2008. GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. En *CLEF*, volumen 5706 de *Lecture Notes in Computer Science*, páginas 808–821. Springer.

Perea-Ortega, José M., Miguel A. García-Cumbreras, Manuel García-Vega, y L. Alfonso Ureña-López. 2008a. Comparing several textual information retrieval systems for the geographical information retrieval task. En *NLDB*, volumen 5039 de *Lecture Notes in Computer Science*, páginas 142–147. Springer.

Perea-Ortega, José M., Luis Alfonso Ureña-López, Manuel García-Vega, y Miguel Angel García-Cumbreras. 2008b. Using query reformulation and keywords in the geographic information retrieval task. En *CLEF*, volumen 5706 de *Lecture Notes in Computer Science*, páginas 855–862. Springer.

Sanderson, M. y J. Kohler. 2004. Analyzing geographic queries. En *Proceedings Workshop on Geographical Information Retrieval SIGIR*.

Spink, Amanda, Bernard J. Jansen, y Cenk H. Ozmultu. 2000. Use of query reformulation and relevance feedback by excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4):317–328.

Stokes, Nicola, Yi Li, Alistair Moffat, y Jiawen Rong. 2008. An empirical study of the effects of nlp components on geographic ir performance. *International Journal of Geographical Information Science*, 22(3):247–264.